Variables, 'random' variables, and probability rules

Statistics for Social Scientists I, L4

Bur, GJM

2024-06-25

1 Probability rules

Now that we've gotten our feet wet with some probability and seen that it is not so bad, let's learn the rules formally. I am keeping these pretty concise here and sticking only to what we need to know for our class. *Some of what we are about to learn won't be immediately useful for a few lectures, but it is awkward to split these rules up.* Our *main* goal in this lecture is to now start talking about probability at the *population level*.

1.1 Prelude to probability rules: random variables

To begin, let's get a bit of definition and notation out of the way. First, some definition.

1.1.1 Definition

So far, we have referred to "variables" a bit willy-nilly. In fact, "variables" can mean *three* things.

1. A variable in a function, like the Y and X in $Y = 3X + 4$.

   Here "variable" simply means "something we can change". We occasionally need to use this meaning, as we did last week in the proof of the mean's property of minimizing the sum of squared deviations.

2. In a statistics context, and with no more modification, the word "variable", as in "the variables in the data-set", means usually something like "this list of data I have", some set of *realizations of a property* (properties like age or income), or some set of *outcomes of an experiment* (e.g. the flip of a fair coin or a subject's choice to purchase something in an online store).

   After all, it *is* fair to call these vectors of information "variables" because they *do* vary from individual to individual. In another sense, however, they are fixed: they can't change as they've already happened, so in that sense they are not really "variable" at all. People's scores on this variable *vary*, but everyone's score is known.

3. Also in a statistics context, a word like "variable" is sometimes prefixed with the word "random". Then, it is used to mean some property of the world that not only can vary from person to person or trial to trial but *is* unknown, *is* just an abstract property.

A variable in this sense is something like the abstract concepts "the outcome of a coin flip" or "a person's income". In this guise, the random variable is a kind of constantly-spinning flipped coin, or constantly-shuffling deck of cards with the names of people to call for a survey.[1] We can take *realizations* of this variable, but the variable is itself not a number or anything finite. In that sense, it is like the letters in the equation $Y = f(X)$. We can plug in values for X, but X itself is a variable, not a number. When we talk about a variable in *this* sense, we should use the phrase "random variable". That term has a more precise definition that we won't get into, but this is the sense in which a variable can be "random".

Summarizing, when we have *sample data* and we talk about "variables", we mean "known scores on some outcome that vary between people". These "variables" are tangible things, column, vectors, lists, etc—however you want to conceive of your data. When we are thinking about our *population* and thinking about the abstract concept of the outcome of a draw from it, like the showing face of a die or the height of a person we pull into our sample, we are using the concept of a random variable.

1.1.2 Notation

We will generally write random variables with *capital* Roman letters, like Y or X. We could also write the names of the variables themselves, like *educ.* or *race* or *inc.*, but this is less common, especially when we are just talking about some generic variable. We usually represent the *generic* outcome of a random variable as a lowercase letter, like y or x. This can be a little confusing because an outcome is a specific instance of something happening, but we are representing it generically. An expression such as "the probability that $Y = y$" can be confusing, but it just means "the event that a random variable takes on some particular value". The usefulness of an expression like this will become clearer soon. Note that the notation itself varies, and this pickiness about notation is new; older textbooks are pretty willy-nilly. *In cases where we want to refer to multiple outcomes of X, it is common to see subscripts like x_1, x_2... etc. This is fine, and we will use it, but just note that we have also already used these to refer also to actual observations in the data, not possible outcomes.* Context will tell you which is which.

1.2 Probability rules!

We first need to review some basic set notation and then revisit Venn diagrams. Does it feel a bit like I'm trying to convince you that every single little "odd and end" that your teacher ever taught you in high school after you had finished the "real" content for the year is … *actual math*? Well, I'm not trying to, but this set of notes will indeed make use of almost every single special topic from high school math about which you wondered "when will I ever use this stuff?"

1.2.1 Set notation

If one outcome of a variable happening is denoted $X = x$, another outcome of another variable $Y = y$, then both-happening is called the intersection of the two events and written $X = x \cap Y = y$. For example, if I roll a six-sided die, and I define one variable X to be "the

[1] Thanks to my friend, colleague, and former TA David Skalinder for giving me this metaphor. I believe he got it from David Kaplan at the University of Wisconsin.

face up is even" and another variable Y to be "the face up is greater than 4", each variable has two possible outcomes. Here is a table showing all the possibilities. Each cell is some specific realization of $X = x \cap Y = y$.

	face > 4	**face** ≤ 4
face even	6	$2, 4$
face odd	5	$1, 3$

Then, the *union* of two outcomes, $X = x \cup Y = y$, means that one outcome happens, the other does, or both do. In the table below, the cells of the table are unions, not intersections.

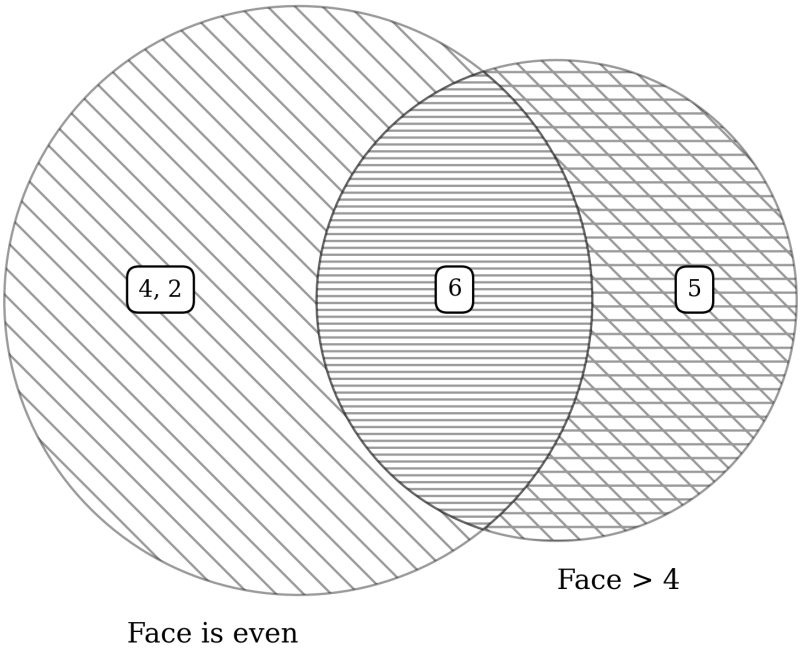
	face > 4	**face** ≤ 4
face even	$2, 4, 5, 6$	$1, 2, 3, 4, 6$
face odd	$1, 3, 5, 6$	$1, 2, 3, 4, 5$

1.2.2 Venn diagrams

We can represent one outcome of each variable as a set on the Venn diagrams below. Here, I chose both outcomes to be "successes" (given how the variables were defined). The intersection represents the case where we have success on both variables; the union, i.e. the set of outcomes in one circle, the other, or both, represents a success on at least one variable, maybe both.[2]

[2]Thanks to Ardell Taugher of SOC360su24 for catching a key mistake here where I accidentally swapped the position of the 5 and the 6 with a Python error.

Venn diagrams for outcomes of rolling a die



Two very important remarks before going forward.

1. Note that I have defined *two* random variables on the same basic *event*, the roll of a die.

   Our random variables could be defined on two different events, where event has the normal English meaning of "something happening in the real world", but in many situations we'll be more interested in random variables *defined on the same event*, including sociology where the single observation of a given person is an "event" which gives rise to many variables defined on that one observation (their income, age, race, height, sex, etc.). *It is more common to ask questions about the probability that someone is poor and religious than to ask questions about the probability that person* 1 *is religious and person* 2 *is poor (in fact, with independent sampling, this question is basically uninteresting).* **The intersections we deal with are usually this type of intersection, where two variables that can intersect are defined on the same event.**

2. We *could*, and sometimes will, calculate the union of outcomes of different variables—if X is "is married" and Y is "miles run weekly", we might look for subjects who satisfy $X = \text{unmarried} \cup Y = 15$. **However, more common, easier to deal with, and what we'll begin with are the unions *of multiple, disjoint outcomes of one variable*, things like "$X =$ divorced $\cup X =$ separated or"$Y \geq 15 \cup Y \leq 5$".** So the table above is not how we'll usually think of unions.

1.2.3 The probability rules

Finally, here are the rules. In what follows, I use $\mathbb{P}[\cdot]$ to mean the probability of some event. This notation isn't universal, but it's common and helps call attention to the fact that probability is a very special thing.

1. **The probability of the union of two events $\mathbb{P}[X = x \cup Y = y] = \mathbb{P}[x] + \mathbb{P}[y] - \mathbb{P}[x \cap y]$.[3] Note that if two events are *disjoint*, meaning that there is no way for both to happen at once (and thus no intersection), we just sum the probabilities.**

   For example, in our table above, the probability of an even number *or* a number larger than 4 is the probability of each outcome summed less their intersection: $\frac{1}{2}+\frac{1}{3}-\frac{1}{6} = \frac{2}{3}$.[4] On our Venn diagram, this is represented by the fact that to find the area of the whole figure, we would add the area of the two circles but, to get rid of double-counting, we'd subtract the area of the intersection once.

   Now, let's think about a more sociological example where we are interested in two *disjoint* outcomes of the *same* variable. What is the probability, according to our religious attendance table above, that someone attends *exactly* weekly or *exactly* monthly? In other words, letting Y represent attendance, what is $\mathbb{P}[Y = \text{weekly} \cup \mathbb{P}[Y = \text{monthly}]$? Since no one can answer both ways, we just write $0.17 + 0.06 = 0.23$. What would this look on a Venn diagram? We'd just have *disjoint* or non-overlapping circles! (Illustration omitted for space).

2. **The probability of one event $Y = y$ happening *given* that $X = x$ has happened is called the conditional probability of Y given X and written $\mathbb{P}[Y|X]$. It is equal to $\frac{\mathbb{P}[X=x \cap Y=y]}{\mathbb{P}[X]}$. Note that if the two events are independent, $\mathbb{P}[Y|X] = \mathbb{P}[Y]$ by definition and we just multiply the probabilities.**

   For example, let's find the probability of getting a number larger than 4 given that we have an even number. First, we find the probability of an even number *and* a number larger than 4, $\frac{1}{6}$, Now, we divide by the condition, which is that the number is even. This is just $\frac{1/6}{1/2} = \frac{1}{3}$. Intuitively, we could also have just listed the even numbers and counted those larger than 4 and gotten the same answer. Note that in this case, our events *are* independent since the overall probability of getting a number larger than 4 was $\frac{1}{3}$ to begin with.

   Let's look at a case where our events *aren't* independent. Let W be a random variable indicating "face of a fair die is divisible evenly by 6". Then, what is $\mathbb{P}[W = 1|X = 1]$? In words, what is the probability that we have a multiple of 6 given that our face is a multiple of 2? We calculate $\frac{\mathbb{P}[W=1 \cap X=1]}{\mathbb{P}[X=1]} = \frac{1/6}{1/2} = \frac{1}{3}$. In other words, although our initial probability of having a multiple of 6 is low, it gets higher once we know that we have a multiple of 2 on our hands. Again, you could also brute-force this by

[3] This rule generalizes for the case of three or more events, but in a way that is a little tricky; the general principle is called the Principle of Inclusion-Exclusion, but we won't need it.

[4] *Note carefully here that each of our variables is just a binary variable*: we either observe a success or a failure. So, the correct notation is not something like $\mathbb{P}[X = \text{even} \cup Y > 4]$ since we already put that information in the variable definitions. Instead, we want $\mathbb{P}[X = \textsc{true} \cup Y = \textsc{true}]$. We'll later see that it is convenient to just represent the outcomes of binary variables, even if they are not themselves truly quantitative, with 0 and 1 representing success and failure.

just restricting your set of outcomes to $2, 4, 6$ and then noticing that one of the three outcomes is a 6.[5]

Let's try a more sociological example. Now, I need to put up what is called a **two-way table**. This gives us another taste of bivariate analysis, which I'll punt on until a later lecture. But, we are slowly easing into it. Below is our table of religious attendance, cross-tabulated with sex. Here, I regrouped the data into monthly vs. less than monthly.

sex	at least monthly	less than monthly	All
male	365	682	1047
female	575	710	1285
All	940	1392	2332

Try calculating the conditional probabilities $\mathbb{P}[\text{attends} = \text{at least monthly, less than monthly}|\text{sex} = \text{male, female}]$ by calculating percentages *across each row* and comparing them to the overall probability of attending at least monthly, $\frac{940}{2332} \approx 0.4$ or not doing so (by the complement rule, 0.6. *Notice that this is a **direct** way of conditioning on some event: we just select rows where the person is male or female.* For example, for men, to find the probability that they attend at least monthly, find $\frac{365}{1047} = 0.35$. This represents the probability of attending monthly *given* that someone is male—and it's noticeably lower!

This *is* equivalent to using the definition we used above. The probability that someone is male and attends services monthly or more is $\frac{365}{2332} \approx 0.157$; the probability that someone is male is $\frac{1047}{2332} \approx 0.449$, so $\frac{0.157}{0.449} \approx 0.35$. Both methods work!

When you fill in the table, you should get the following.

sex	at least monthly	less than monthly
male	0.349	0.651
female	0.447	0.553
All	0.403	0.597

Note that this is a form of rudimentary bivariate analysis!

3. **The probability that an event happens or anything *but* it happens is 1.** For simplicity, suppose all events in the world are considered outcomes of some variable X. Then, $\mathbb{P}[X = x \cup X = \neg x] = 1$. This is a consequence of the first rule since x and

[5] By the way, how do we represent this property of dependence/independence on a Venn diagram? It's a little hard to see, but if you imagine the background of the circles as being the whole outcome space (everything that could happen), if the ratio of the intersection of events to one of the individual circles is the same as the ratio of that circle to the total space, events are independent. Bizarrely, the book for this course, which is about as good as it gets for this level but not phenomenal, claims that independence can't be seen on a Venn diagram: "Unlike disjointness, we cannot depict independence in a Venn diagram because it involves the probabilities of the events rather than just the outcomes that make up the events". This is false.

$\neg x$ are disjoint. This rule is also called the *complement rule*. It is often easier to work out the probability of $X = x$ by finding $1 - \mathbb{P}[X = \neg x]$.

Here is a fun, classic example. If you have 26 randomly-selected people in a room, what is the probability that at least two share the same birthday? Assume that birthdays are uniformly distributed over the year, which is false but not wildly so. It is hard to work out all the probabilities of high-order intersections of birthday-sharing (e.g., what's the probability that 17 people share one? This is actually a really good thing for us to know how to calculate at higher levels, but it isn't easy). It is, however, easy to work out the probability that none of them share a birthday.

We do so by reasoning that the probability that with two people, the probability that they do *not* share one is $\frac{364}{365}$ since any other day of the year besides person 1's birthday would make this true (and we assume all other days are equally probable), for three people $\frac{364 \cdot 363}{365^2}$... and so on. If we actually calculate this out, we have $\frac{1}{365^{25}} \cdot \frac{364!}{339!} = 0.40$, meaning that there is a 60 percent probability that two or more people do share a birthday in a room of only 26 people. This is a pretty surprising fact!

4. **The probability of some event $\mathbb{P}[Y = y|X = x]$,can be related to the *inverse* conditional probability, $\mathbb{P}[X = x|Y = y]$, in the following way, which is often useful if we only have information on the inverse *or* if people are prone to confusing the two.**

$$\mathbb{P}[Y = y|X = x] = \frac{1}{\mathbb{P}[Y = y]}\mathbb{P}[X = x|Y = y] \cdot \mathbb{P}[Y = y]$$

We will study this more later when we talk about relationships between variables. It is called **Bayes' theorem**.

2 Population statistics

Next lecture, we'll put these probability facts to good use in exploring density curves, an important use which we have not yet explored. Now, let's make a quick stop at *population statistics* that are analogous to the sample statistics we've already developed.

We've been saying that some properties in our sample are estimates of the corresponding population properties. For example, our sample mean $\overline{y}$ estimates our population mean μ_Y; our sample variance s^2 estimates our population variance σ^2; our sample PMF of a qualitative variable estimates our population PMF.

However, we haven't actually said how our sample *means and variances* are calculated in the population. We had actual *formulae* for $\overline{y}$ and s_y^2, namely $\frac{1}{n}\sum_{j=1}^{n} y_j$ and $\frac{1}{n-1}\sum_{j=1}^{n}(y_j - \overline{y})^2$, respectively. What are the formulae in the population?

2.1 Population mean: expectation or expected value

The population mean is called the **expectation** or **expected value** (for us, synonymous). For a *discrete* quantitative random variable, it is simply the weighted mean of possible outcomes.

We write the expected value *formula* as $\mathbb{E}[Y]$, which is analogous to the summation formula in the sample.[6] The "quick notation" in the population is μ_Y, analogous to $\overline{y}$; μ_Y is like $\overline{y}$ in that it just means "the mean" with no reference to how it's calculated. So, how is it calculated? The exact definition is as follows.

$$\mu_Y = \mathbb{E}[Y] = \sum_y \mathbb{P}[Y = y] \cdot y$$

You might wonder what exactly this means. The notation can be intimidating, but the idea is simple. The sum over "little y" simply means summing over all possible outcomes of the variable. Didn't we sum over *individual observations* in all of our previous summations? Yes, we did. However, when we work at the population level, we don't always have a finite number of trials to work with. How many rolls of a fair die are there? An impossible question! And after all, we can sum over any index that makes sense.

Let's try calculating an example quickly, before we get too lost in the minutiae. What is the expected value of the roll of a fair die, anyways?

$$\begin{aligned}
\mathbb{E}[\text{die-face}] &= \sum_{y=1}^{6} \frac{1}{6} y \\
&= \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) \\
&= \frac{1}{6}\frac{6 \cdot 7}{2} \\
&= \frac{1}{6}\frac{42}{2} \\
&= 3.5
\end{aligned}$$

In the fourth line, I used the fact that the sum of the first n positive integers is $\frac{n(n+1)}{2}$, a fact that we'll actually have occasion to prove later. You could also just count it up by hand, of course.

OK, so what does this actually *mean*? On one hand, it's simply the average face-value of the die. One natural way to think about the expectation is the price you would be willing to pay to gamble. For example, in our die roll situation, suppose that you are going to have a single try at the following game: I will roll a die and pay you $100y$ dollars, where y is the face-value of the die. A good deal! But, what if I charge you 200 to play—is it such a good deal then? The answer, at least in the long run, is actually *yes*, since the expected winnings for you would be 350; whether you would actually be rational to play this game if you only have one turn is a messier question.

Let's try another example, which in fact simpler and more useful. What is the expectation of a biased coin if we write down a score of 1 if we get heads and for tails a 0, if the coin is biased with probability p towards a heads? Let's call this variable B for binary.

[6] The fancy "blackboard bold" E just represents the fact that this is a very important operator (sometimes notation rules aren't very exact).

$$\begin{aligned}\mathbb{E}[B] &= \sum_{y=1}^{2} \mathbb{P}[Y = y]y \\ &= p(1) + (1-p)0 \\ &= p\end{aligned}$$

In other words, the expectation of a binary variable which is given values 0 and 1 for `FALSE` and `TRUE` respectively is just the probability of success. This will be very useful later. It suggests that if we *do* have a binary qualitative variable, we should code it this way for ease of interpretation.

2.1.1 Expectation as a mean

In case it wasn't clear above, the expectation is really not anything but a mean. While the expectation is *theoretically* prior to the simple arithmetic mean, it is almost certain to have come after it historically since the expected value dates only to the early modern period. We can think about the expectation as simply generalizing the arithmetic mean. Let's recall the formula for the "empirical" mean:

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$$

Now, if there are repeated values in our sample, say n_h copies of some value y_h, the contribution of those n_h people to the total is simply $n_h \cdot y_h$. So, we could actually just find our sample mean by summing over unique values of the variable—just like we were doing before with the expectation. Let's write that out. Let h just symbolize our unique values, and suppose that our population *is*, in fact, finite. That won't always be the case, but I want to show that *if* the population is finite, the "sample" mean formula and the "population" expectation formula are the same.

$$\begin{aligned}\bar{y} &= \frac{1}{n}\sum_{h} n_h y_h \\ &= \sum_{h} \frac{n_h}{n} y_h \\ &= \sum_{h} \mathbb{P}[Y = y_h]y_h \\ &= \mathbb{E}[Y]\end{aligned}$$

Since the sample mean can be written as the probability of any given value multiplied by that value, we often write the expectation as μ. This is more-or-less the same thing as writing $\overline{y}$ in place of the expression $\frac{1}{n}\sum_{j=1}^{n} y_j$. In what follows, I will usually refer to the mean as μ, subscripting it where necessary to avoid confusion. The reason for the $\hat{\mu}$ notation for the sample mean is that the "hats" (technically, carets) generally mean "estimated". This notation is logical, but it is less common than the (inconsistent) bar-notation.

2.1.2 Expectation properties

I won't prove these in the main text, but the idea is that all of our major properties of the mean also apply to the expectation. They are generally pretty easy to prove because an expectation is just a summation. Here they are quickly.

1. $\mathbb{E}[kY] = k\mathbb{E}[Y]$
2. $\mathbb{E}[Y + k] = \mathbb{E}[Y] + k$
3. $\mathbb{E}[Y + X] = \mathbb{E}[Y] + \mathbb{E}[X]$

2.2 Variance

The *variance*, too, can be expressed as a population level mean. We write it as σ^2 (note that this is a *lowercase* sigma, for **s**tandard deviation). Again, it would be more logical to call its sample counterpart $\hat{\sigma}^2$, but we are stuck with s^2. We write out the variance as an expectation at the population level.

$$\sigma_Y^2 = \mathbb{V}[Y] = \mathbb{E}[(Y - \mu_Y)^2]$$

I'll show just one quick property I mentioned last week, the Köning-Huygens identity. It is *exactly*, not just approximately, correct at the population level. It makes calculating variances *very* easy when working with population data about which we know a little something.

$$\begin{aligned}
\mathbb{V}[Y] &= \mathbb{E}[(Y - \mu_Y)^2] \\
&= \mathbb{E}[Y^2 - 2Y\mu_Y + \mu_Y^2] \\
&= \mathbb{E}[Y^2] - 2\mu_Y\mathbb{E}[Y] + \mathbb{E}[\mu_Y^2] \\
&= \mathbb{E}[Y^2] - 2\mathbb{E}[Y]^2 + \mathbb{E}[Y]^2 \\
&= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2
\end{aligned}$$

Using this formula, let's calculate the variance of our coin flip and die roll, mentioned above. For the die roll, we have…

$$\begin{aligned}
\mathbb{V}[\text{die-face}] &= \mathbb{E}[D^2] - \mathbb{E}[D]^2 \\
&= \frac{1}{6}\sum_{h=1}^{6} h^2 - 3.5^2 \\
&= \frac{1}{6}(1 + 4 + 9 + 16 + 25 + 36) - 3.5^2 \\
&= \frac{91}{6} - \frac{49}{4} \\
&= 2.92
\end{aligned}$$

There is a formula for the sum of squared integers, by the way, that we can use to quickly get 91, but I won't prove it here. See L1 appendix.

Let's again work an example using the K-H formula and the coin flip now.

$$\begin{aligned}\mathbb{V}[B] &= \mathbb{E}[B^2] - \mathbb{E}[B]^2 \\ &= p(1^2) + (1-p)(0^2) - p^2 \\ &= p - p^2 \\ &= p(1-p)\end{aligned}$$

This is also a very nice, simple formula—we should *definitely* take advantage of this and code our binary qualitative variables this way!

2.2.1 Variance properties

The following two properties, both analogous to our sample properties, hold.

1. $\mathbb{V}[kY] = k^2\mathbb{V}[Y]$
2. $\mathbb{V}[Y + k] = \mathbb{V}[Y]$

3 Exercises

No additional execises for this lecture.

4 Appendix: basic population statistics

Again, this is not necessary for our class. It's just here for reference. But you may find this useful at higher levels!

4.1 Population expectation

Let's define the **expectation** or **expected value** (for us, totally synonymous) for a discrete random variable as the probability-weighted sum of possible outcomes of a random variable. Recall that "random" here is misleading; the *randomness* comes from the fact that any realization of the variable is random conditional on the probabilities of the outcomes (but the normal meaning of randomness would imply that these probabilities are *uniform*, which does not *need* to be the case, by any means).

4.1.1 Expectation as a mean

While the expectation is *theoretically* prior to the simple arithmetic mean, it is almost certain to have come after it historically since the expected value dates only to the early modern period. We can think about the expectation as generalizing the arithmetic mean. We can write the formula for the "empirical" mean as follows.

$$\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$$

Now, if there are any repeated values, it is intuitive that we could rewrite the mean as follows, where $y \in \Omega$ indexes the unique possible outcomes of the variable and n_y is the number of occurrences of outcome $Y = y$. Note that summing over y means simply to sum over all unique values.

$$\begin{aligned}\bar{y} &= \sum_y n_y \frac{y}{n} \\ &= \sum_y \frac{n_y}{n} y \\ &= \sum_y \mathbb{P}[Y = y] y\end{aligned}$$

As we can see, the sample mean can be written as the probability of any given value multiplied by that value.

So, we define the expectation as follows.

$$\mathbb{E}[Y] = \sum_y \mathbb{P}[Y = y] y$$

4.1.2 Expectation properties

The expectation is linear: $\mathbb{E}[aY + b] = a\mathbb{E}[Y] + b$. We show this in two parts.

$$\begin{aligned}\mathbb{E}[aY] &= \sum_y \mathbb{P}[Y = y] ay \\ &= a \sum_y \mathbb{P}[Y = y] y \\ &= a\mathbb{E}[Y]\end{aligned}$$

Now, we show that …

$$\begin{aligned}\mathbb{E}[Y + b] &= \sum_y \mathbb{P}[Y = y](y + b) \\ &= \sum_y \mathbb{P}[Y = y] y + \sum_y \mathbb{P}[Y = y] b \\ &= \mathbb{E}[Y] + b \sum_y \mathbb{P}[Y = y] \\ &= \mathbb{E}[Y] + b\end{aligned}$$

In the penultimate line, we used the fact that b can be factored out of the summation since it does not depend on the value of y. Then, the sum over the probabilities of y is simply 1. This property can be generalized: the expectation of a constant is just that constant.

And the expectation of two random variables is as follows. I'll switch to the notation $\pi_{x,y}$ to mean $\mathbb{P}[X = x \wedge Y = y]$, which is more concise. Also, note that summing over lowercase x and y means summing over all possible realizations of the variables X and Y.

$$
\begin{aligned}
\mathbb{E}[X+Y] &= \sum_{x \wedge y} \pi_{x,y}(x+y) \\
&= \sum_x \sum_y \pi_{x,y}(x+y) \\
&= \sum_x \sum_y \pi_{x,y} y + \sum_x \sum_y \pi_{x,y} x && \text{distributive property} \\
&= \sum_y y \sum_x \pi_{x,y} + \sum_x x \sum_y \pi_{x,y} && \text{changing the order of summation, moving out variables that don't deper} \\
&= \sum_y y \pi_y + \sum_x x \pi_x && \text{see below for full explanation} \\
&= \mathbb{E}[Y] + \mathbb{E}[X] && \text{definition of expectation}
\end{aligned}
$$

In the penultimate line, we used the fact that summing the joint probability $\mathbb{P}[X = x \wedge Y = y]$ over, say, x alone means to sum over every possible outcome of X. For a fixed $Y = y$, this joint probability is then simply equivalent to the probability $Y = y$. For example, consider the probability of rolling a regular six-sided die (represented by X) and flipping a coin (represented by Y) at the same time. For either outcome of the coin-flip $y = 0, y = 1$, summing up the joint probabilities $X = 1 \wedge Y = y, X = 2 \wedge Y = y ... X = 6 \wedge Y = y$ must give the probability of flipping that side of the coin.

The intuition here is that the expectation of a sum of random variables is literally equal to the sum of products of the joint probability of any pair x, y and the sum $x + y$. We can break that up into the sum of products of the joint probability of any pair x, y and each variable's realization, *x or y*. Then, both of those terms can be regarded not as a double summation over each each variable (a substitution I made in step two above): summing over all pairs x, y can be written as a single or a double summation with appropriate indexing. Finally, summing the joint probability over either variable individual just leave the marginal probability of the other at some value of it. Then, summing the product of the marginal probabilities of a variable's realizations with the value of the realization simply *is* the expected value.

4.1.3 Conditional expectations

The conditional expectation is simply the expectation of some variable Y given the value of some variable X; it is written $\mathbb{E}[Y|X]$. This conditional expectation is itself a random variable; while the expectation of a random variable is a scalar, the conditional expectation depends on the value of a random variable.

$$\mathbb{E}[Y|X] = \sum_y y \cdot \mathbb{P}[Y = y|X]$$

4.1.3.1 Law of total expectations (AKA LTE or law of iterated expectations)

The conditional expectation also has an expectation, and this gives us a very convenient fact known as the law of iterated expectations or law of total expectation: $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$. The proof is quite intuitive: after all, the expected value of all possible conditional expectations is basically just the expected value taken over all situations, and this should logically be the simple expectation. But, we can also prove this algebraically.

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[Y|X]] &= \mathbb{E}\left\{\sum_y y \cdot \mathbb{P}[Y = y|X]\right\} \\
&= \mathbb{E}\left\{\sum_y y \cdot \frac{\mathbb{P}[Y = y \wedge X = x]}{\mathbb{P}[X = x]}\right\} && \text{definition of conditional probability} \\
&= \sum_x \sum_y y \cdot \frac{\mathbb{P}[Y = y \wedge X = x]}{\mathbb{P}[X = x]} \mathbb{P}[X = x] && \text{definition of expectation} \\
&= \sum_x \sum_y y \cdot \mathbb{P}[Y = y \wedge X = x] && \text{algebra} \\
&= \sum_y y \cdot \mathbb{P}[Y = y] && \text{discussed above} \\
&= \mathbb{E}[Y] && \text{definition of an expectation}
\end{aligned}$$

4.2 Population variance

The *variance* is another basic property of a distribution or summary statistic. It is one way of operationalizing the more general concept of *variation* or *spread*. The variance is specifically defined as the expected *squared* deviation of a variable from its mean.

$$\mathbb{V}[Y] = \mathbb{E}[(Y - \mu_Y)^2]$$

Although it is not common to actually see it written out as an expectation, it is sometimes useful.

$$\begin{aligned}
\mathbb{V}[Y] &= \mathbb{E}[(Y - \mu_Y)^2] \\
&= \sum_y \mathbb{P}[Y = y](y - \mu_Y)^2
\end{aligned}$$

4.2.1 The Köning-Huygens formula for the variance

We can simplify this expression in a way very common in advanced applications, sometimes called the Köning-Huygens identity.

$$\begin{aligned}
\mathbb{V}[Y] &= \mathbb{E}[(Y - \mu_Y)^2] \\
&= \mathbb{E}[Y^2 - 2Y\mu_Y + \mu_Y^2] \\
&= \mathbb{E}[Y^2] - 2\mu_Y\mathbb{E}[Y] + \mathbb{E}[\mu_Y^2] \\
&= \mathbb{E}[Y^2] - 2\mathbb{E}[Y]^2 + \mathbb{E}[Y]^2 \\
&= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2
\end{aligned}$$

Its prominence is because the *square* is more mathematically tractable than, say, the absolute value of a variable's mean deviation. We use the square of the deviation because the expectation of a random variable about its mean is zero.

$$\mathbb{E}[(Y - \mu_Y)] = \mu_Y - \mu_Y = 0$$

4.3 Conditional variance

We should also define the *conditional variance*. This can can be a little tricky. The variance of a random variable is a parameter or a moment: once we choose some actual variable to assume the place of Y, the variance is just a number (assuming we know its probability distribution; even if we don't, we can estimate it with a sample).

The conditional variance is, however, itself a random variable. Conceptually, it is the variance of some variable of interest Y evaluated at some value of a variable that is (hopefully) related to it, which we'll denote X. Given a *value* of X, some x, we could evaluate the conditional variance at that value, and one possible confusion here is that in an actual sample, the phrase "conditional variance" usually refers to some specific conditional variance. But without specifying that $X = x$, we just have a random variable. Below, I will use the notation $\mu_{Y|X}$ to indicate the conditional mean, which isn't extremely common but which is much easier to look at for me (in general, while expectation notation and parameter notation imply different things in some contexts, they often mean more-or-less the same thing).

$$\begin{aligned}
\mathbb{V}[Y|X] &= \mathbb{E}[(Y - \mathbb{E}[Y|X])^2|X] \\
&= \mathbb{E}[(Y - \mu_{Y|X})^2|X] \\
&= \sum_y \mathbb{P}[Y = y|X](y - \mu_{Y|X})^2
\end{aligned}$$

Note that in the final line, in an actual sample, we calculate a conditional variance by only using observations in the group or condition. At the population level, however, we presume that we know the *relative* distribution of values of Y given X. So, the analogue here to "restricting our sample to only those observations with value $X = x$" is calculating the conditional probabilities $\mathbb{P}[(Y = y)|X]$ rather than the marginal probabilities $\mathbb{P}[(Y = y)]$.

Note also that we can also just write down the K-H identity and then substitute in carefully. The conditional operator is very general and can just mean "evaluated at", so we plug in $Y|X$ wherever we see Y (in the first expression, the expectation of the square-of-Y given X is the same as the expectation of the square of Y-given-X).

$$\mathbb{V}[Y|X] = \mathbb{E}[(Y^2|X)] - \mathbb{E}[Y|X]^2$$

4.4 The "taking out what is known" property

Finally, we also need the "taking out what is known" property of expectations of the sort $\mathbb{E}[f(X)Y|X]$. Since $f(X)$ evaluated at any particular $X = x$ is just a number that does not depend on Y, it is can be moved outside of the summation over the conditional expectation.

$$\begin{aligned}
\mathbb{E}[f(X)Y|X] &= \sum_y \mathbb{P}[Y = y|X = x]f(X)y \\
&= f(X)\sum_y \mathbb{P}[Y = y|X = x]y \\
&= f(X)\mathbb{E}[Y|X]
\end{aligned}$$

This is especially important for the law of total expectation.

$$\begin{aligned}
\mathbb{E}_Y[f(X)Y] &= \mathbb{E}_X[\mathbb{E}_Y[f(X)Y|X]] \\
&= \mathbb{E}_X[f(X)\mathbb{E}_Y[Y|X]]
\end{aligned}$$

4.5 Law of total variance

Then, finally, the variance of Y can be written as follows; this is called the "law of total variance". In the following, I will in some places use $\mu_{Y|X}$ as shorthand for the expectation of Y given X for the sake of concision. I also will emphasize that the expectation we initially took was over Y by subscripting the other expectations where convenient.

$$\begin{aligned}
\mathbb{V}[Y] &= \mathbb{E}[(Y - \mu_Y)^2] \\
&= \mathbb{E}[(Y - \mu_{Y|X} + \mu_{Y|X} - \mu_Y)^2] && \text{Add/subtract const.} \\
&= \mathbb{E}_Y[(Y - \mu_{Y|X})^2] + \mathbb{E}_Y[(\mu_{Y|X} - \mu_Y)^2] + 2\mathbb{E}_Y[(Y - \mu_{Y|X})(\mu_{Y|X} - \mu_Y)] && \text{Bloc binomial} \\
&= \mathbb{E}_X\left\{\mathbb{E}_Y[(Y - \mu_{Y|X})^2|X]\right\} + \mathbb{E}_X[(\mu_{Y|X} - \mu_Y)^2] + \ldots 0 && \text{LTE; see below} \\
&= \underbrace{\mathbb{E}_X\left\{\mathbb{E}_Y[(Y - \mu_{Y|X})^2|X]\right\}}_{\mathbb{E}[\mathbb{V}[Y|X]]} + \underbrace{\mathbb{E}_X[(\mu_{Y|X} - \mu_Y)^2]}_{\mathbb{V}[\mathbb{E}[Y|X]]} && \text{See below}
\end{aligned}$$

In the penultimate line, two things happen that merit more comment.

First, since the second term does not actually depend on Y at all beyond $\mu_{Y|X}$, we can take the expectation over X and then remove the expectation over Y; further, the conditional part of the expectation can be eliminated since the only random variable involved is X.

Second, the cross-product term alluded ends up being zero. Here, I will show how. We can ignore the leading constant 2 for simplicity. Then, we factor out the terms that do not actually depend on Y.

$$\begin{aligned}&(\mu_{Y|X} - \mu_Y)\mathbb{E}_Y[(Y - \mu_{Y|X})] \\ =&\mu_{Y|X}\mathbb{E}_Y[(Y - \mu_{Y|X})] - \mu_Y\mathbb{E}_Y[(Y - \mu_{Y|X})]\end{aligned}$$

Now, if we take the expectation over X, in line with the LTE, we have…

$$\mathbb{E}_X\left\{\mu_{Y|X}\mathbb{E}_Y[(Y - \mu_{Y|X})]|X\right\} - \mathbb{E}_X\left\{\mu_Y\mathbb{E}_Y[(Y - \mu_{Y|X})]|X\right\}$$

Now, both of these terms are zero. The second term is easier; we could actually factor the μ_Y out of the whole thing and just have the conditional expectation of the variable Y minus…the very same thing. The first term can be dealt with in one of two ways. First, we could simply expand and take the conditional expectation in parts; this is clear, if tedious. But, we could also just realize, conceptually, that the conditional expectation of the deviation of Y at $X = x$ from the conditional mean of Y at $X = x$ is, simply, the same thing, said two different ways. *Even though we have technically the product of two random variables inside the larger expectation*, typically a problem, we can basically ignore that here: multiplying by the conditional expectation of Y given X, whatever it is, cannot stop the deviation of that expectation from itself being zero.